Local vs. global scope of discourse markers

Local vs. Global Scope of Discourse Markers: Corpus-based Evidence from Syntax and Co-occurring Pauses

Ludivine Crible

University of Louvain-la-Neuve

Abstract

This paper discusses the relevance and challenges of a corpus-based investigation of the scope of discourse markers for cognitive semantics. It builds on Lenk's (1998) distinction between local and global scope of discourse markers and strives to map it with annotation variables available in existing corpora (i.e. extent and location of arguments). Given the interplay of syntactic and semantic-pragmatic variables that a direct approach to scope involves, it is argued that indirect and independent cues (namely position of the marker, its degree of syntactic integration and co-occurrence with pauses) offer a more reliable access to the variation in scope. The analysis focuses on three pairs of discourse markers and their annotation in a comparable corpus of spoken English and French.

Key words: discourse markers, scope, pauses, position, annotation, corpus-based, cognitive semantics

1. Introduction

Discourse coherence in spoken language is constrained by temporal dynamics imposing the urgency and pressure of the present while maintaining connections with the previous context, or "retentions", and setting the scene for upcoming material, or "projections" (Deppermann and Günthner 2015). These backward- and forward-looking operations can affect various levels of language structure, from local syntax (verbal dependency relations) to global discourse (co-reference, coherence). Some linguistic devices are particularly suited to signal these non-linear connections: chief among them, the category of discourse markers (henceforth DMs, Schiffrin 1987) is dedicated to the management of "local and global content and structure" (Fischer 2000: 20) through a very broad functional spectrum fulfilled by heterogeneous expressions such as conjunctions (*and*, *so*, *although*), adverbs (*actually*, *well*) but also verb phrases (*I mean*, *you know*) or interjections (*yeah*, *oh*), among others.

Studies of discourse markers (or connectives) in written language tend to view them as cohesive ties building up a rather shallow discourse structure as signals of causal or contrastive relations, for instance: this line of research is primarily represented by the very influential Penn Discourse Treebank 2.0

(henceforth PDTB, Prasad et al. 2008) or the Cognitive approach to Coherence Relations (henceforth CCR, Sanders et al. 1992). Confrontation to spoken data, however, soon reveals that the same items (e.g. *so*, *but*) show instantiations of both local (relational) and global (non-relational) uses as signposts to a higher level of discourse organization. As a result, the traditional representation "Arg1-DM-Arg2", where the DM connects two simple and adjacent arguments, is often incompatible with the intricate, non-linear structure of spoken discourse.

This article builds on Lenk's (1998) distinction between local vs. global scope of discourse markers, which she respectively associates with utterance relations (cause, contrast, etc.) and topic relations (topic-shift, topic-resume, etc.) at each end of the continuum. Of course, the divide is not binary and a fine-grained approach to DM scope should also account for intermediate cases where utterance relations are more distant and far-reaching (e.g. a conclusion over multiple utterances) and where topic relations manage shorter segments (e.g. resuming the previous topic after a short single-sentence digression). The absence of one-to-one mapping between specific DMs, their functions and their arguments calls for a more systematic investigation of the notion of scope grounded in empirical evidence, disentangling the interplay of syntactic and pragmatic factors in the behavior of local vs. global DMs.

The feature of DM scope has been addressed rather irregularly in previous corpus-based research where authors often target some (but not all) variables involved in its investigation, including large-scale bottom-up identification of discourse markers, sense disambiguation covering both local-cohesive and global-structuring functions, annotation of their arguments and full discourse segmentation in units of various sizes. In spoken corpora, in particular, such an ambitious undertaking might be even more challenging: [Author] (in press) state that "explicitly identifying the units under a DM's scope may be too ambitious (at least for spoken data)"; they further argue that "sense disambiguation is informative and complex enough and should not necessarily be combined with an identification of the related segments".

The present paper starts from this observation of how challenging (even impossible) a systematic annotation of DM scope would be in spoken corpora and rather provides indirect yet operational cues to the variability of local vs. global functions of DMs, converging evidence from mainly three types of linguistic analysis: i) sense disambiguation of all DMs in a comparable English-French spoken corpus, ii) annotation of position and degree of syntactic integration of DMs and iii) identification of co-occurring pauses. The underlying hypothesis states that pauses are windows to the cognitive processing of local vs. global scope, which should in turn be linguistically reflected by different syntactic (position) and syntagmatic (co-occurrence)

behaviors. This study thus falls within the usage-based framework of cognitive semantics, whereby converging independent evidence of forms and functions is taken as a reliable methodological gateway to "this infamously slippery object of study, semantics" (Glynn 2010: 240). The analysis focuses on three pairs of DMs potentially related to different degrees of scope, namely topic-shift vs. topic-resume (Section 4.1), subordination vs. coordination (Section 4.2) and consecutive vs. conclusive uses of the DM *so* (Section 4.3). Theoretical background and materials will be presented in the following sections.

2. Accessing DM scope through direct and indirect evidence

Discourse markers are here broadly defined as procedural, syntactically optional expressions functioning at discourse-level to "integrate their host utterance into a developing mental model of the discourse in such a way as to make that utterance appear optimally coherent" (Hansen 2006: 25). They constitute a formally heterogeneous class whose functional spectrum covers discourse relations, metadiscursive comments, topic structure and interactional management, following several classification models (González 2005; Cuenca 2013; [Author] 2017a). With such a formal-functional definition in mind, the following sections will develop the notion of DM

scope, its treatment in previous research and its relevance for cognitive linguistics in general and the present study in particular.

2.1 Previous approaches to DM scope

Most definitions of discourse markers agree on the lower boundary of units minimally qualifying for the status of discourse-level argument: an item is only considered to act as a discourse marker if it takes scope over at least a clause(-like) or larger unit (e.g. the "elementary discourse units" in Rhetorical Structure Theory, henceforth RST, Mann and Thompson 1988). There is, however, no principled upper limit as to the extent of arguments under a DM's scope, be it multiple utterances, whole turns or entire interactions. Unger (1996) was one of the first authors to explicitly address the notion of DM scope with respect to the extent of the related units: he acknowledges that "discourse connectives can have scope over an utterance or a group of utterances" (1996: 409), yet admits that "though a paragraph break broadens the range of assumptions serving as candidates for the choice of a context, one particular utterance within a preceding paragraph may still be the most likely candidate" (1996: 436); in other words, a DM introducing a new paragraph does not necessarily take as its Arg1 the full previous paragraph. The identification of a DM's arguments is therefore not a trivial step in the analysis and impacts the functional disambiguation: [Author] (in press)

discuss an example where a particular DM can be assigned different senses depending on the choice of Arg1, and conclude that DMs, especially in speech, tend to "combine local and global scope simultaneously", which makes the annotation process quite challenging.

Yet, annotation of DM scope (in the form of arguments identification) is central in many writing-based frameworks, where the notion is operationalized and systematically annotatedited In the PDTB corpus, for instance, extent (single vs. multiple) and location (adjacent vs. non-adjacent) of the first argument (Arg1) of a given connective are annotated and the results show that 3.34% of all explicit connectives take scope over multiple utterances while, in 9% of the cases, Arg1 is non-adjacent to Arg2. These rather low proportions might be explained by the limited range of DM functions included in the PDTB taxonomy, which does not include any global functions but only allows local discourse relations (e.g. consequence) to be used more globally across multiple and/or distant utterances[1]. Typically global functions include topic relations (topic-shift, topic-resume) or turn-exchange functions (turn-opening, turn-closing) which target hierarchically larger units than utterances. This divide between local and global functions is sometimes conveyed at a terminological level by distinguishing connectives

[1] In the PDTB 2.0, the "list" relation could be considered as potentially global, since elements of an enumeration can be rather distant in a written text. However, in the latest version (PDTB 3, e.g. Webber et al. 2016), this relation type was removed from the taxonomy.

(typically local, cohesive) from discourse markers (typically global, coherent), as in Schiffrin (1987) or Cuenca (2013). However, Lenk (1998) shows that a single item – she focuses on *however* and *still* in spoken British and American English – can express both local and global meanings. This multifunctionality of DMs is also addressed by Bunt (2012) who relates it to the multidimensional nature of dialogs, "involving multiple activities at the same time, such as making progress in a given task or activity; monitoring attention and understanding; taking turns; managing time, and so on" (Bunt 2012: 243). An adequate analysis of DM scope in spoken data should therefore come to terms with the multifunctionality of DMs and account for functions at a higher level of discourse organization (e.g. topic-shift).

One major framework which addresses these aspects of scope is RST and its application to the RST Signalling Corpus (Das et al. 2015) which contains newspaper articles fully annotated for discourse relations (including topic relations) and their signals, distributed over a tree-based segmentation of texts in arguments of different sizes. However, no such undertaking is currently available for spoken corpora, to date: Stent (2000: 250) admits that "given the length and complexity of a typical dialog, it may not be possible to achieve complete coverage", as opposed to written texts where each unit forms a pair with another and each pair is itself hierarchically included in a

higher-order relation until full-text segmentation is achieved[2]. Speech-specific models of discourse segmentation have been proposed: one of them is the Val.Es.Co 2.0 corpus (Cabedo and Pons 2013) where full conversations are segmented hierarchically into more or less local units such as subacts, acts, turns, interventions, etc. In their approach, however, the functions of DMs are only defined at a coarse-grained level, distinguishing among textual, interpersonal and modal types. A more fine-grained study using the Val.Es.Co system is provided by Estellés Arguedas and Pons Bordería (2014) who identified the specific pattern of DMs (e.g. Spanish *bueno* 'well') in "absolute initial position" when signalling a major change in context such as an increase of speakers or a change of speaker status.

In sum, a systematic analysis of DM scope which combines sense disambiguation and argument identification seems to require full discourse segmentation, as in the RST and Val.Es.Co models. However, these tasks are very costly and challenging to implement reliably; in addition, they demand a substantial involvement of the analyst's subjectivity: disambiguating the meaning-in-context of a DM and identifying the arguments in its scope are two strongly inter-related, even circular steps in the analysis, where one decision impacts the other (cf. [Author], in press). Therefore, it might be argued that a cognitive-semantic approach to DM scope should rather turn to

[2] See also Baldridge and Lascarides (2005) on a similar observation of the limitation of Segmented Discourse Representation Theory (SDRT) in dialogs.

more objective, non-circular evidence of the difference between local and global DMs, which can in turn be reliably related to general cognitive theories, as we will now come to see.

2.2 Cognitive correlates to DM scope

DM scope is not only important as an annotation variable in corpus-based research it is also relevant to general processes involved in the cognition of speech. The division of labor between local and global scope of DMs can indeed be related to Levelt's (1989) microplanning vs. macroplanning which are two of the speaker's mental tasks respectively dealing with i) the structure and style of an utterance and ii) designing the communicative intention. Both activities are cognitively demanding, so much so that speakers tend to attend to them separately, in alternation: several experimental studies (Beattie 1980; Greene and Cappella 1986; Roberts and Kirsner 2000) suggest that macroplanning takes place during major pauses preceding a coherent discourse segment (or "cycle") after which temporal fluency (i.e. non-interruption) can be resumed while speakers only attend to microplanning. It would thus seem that micro- and macroplanning are respectively associated to low and high demands on cognitive processing, which is directly observable by longer pauses and more hesitations at the boundary between two cycles of macroplanning.

Although Levelt's (1989) dichotomy did not originally target degrees of DM scope, it can easily accommodate them: local DMs connect or take scope over adjacent units of which they make the linkage explicit, thus managing rhetorical effects (1); global DMs announce more far-reaching connections with distant and/or larger units that constitute major building blocks in the elaboration of the whole discourse structure (2).

(1) *I wasn't looking forward to doing it* but *I am now (EN-phon-01)*[3]

(2)

ICE_10	so *what did you do today then*
ICE_9	*today (0.700) I went I watched the Grand Prix (2.047) and then uh do you remember a neighbour in Hillside called uh the Pembertons*
ICE_10	*yes Pembertons*
ICE_9	*well I know uh (0.770) I met him actually about a year ago with uhm*
ICE_10	*[...] Oliver?*
ICE_9	*yeah Oliver*

[3] All examples in this paper come from the *DisFrEn* corpus ([Author], 2017b), see Section 3 for more details.

ICE_10 *didn't I go to school with their daughter is there is there was there a girl there [...] was there a sister there*

ICE_9 *well uh he's got a (0.330) I don't know whether yeah I suppose so [...] he's got somebody living in his house who used to go to Mrs. Parsons*

ICE_10 so *how did you meet up with him then*

ICE_9 *oh he was a member of the bicycle polo club last year*

ICE_10 *oh right (2.560) what kind of bicycles do you ride on then*

ICE_9 *bicycles with two wheels handlebars and a frame [...] the wheels are very close together so you can turn quickly*

ICE_10 so *where did you play this*

ICE_9 *Uhm in Putney (1.470) Hurlingham Park [...] it's next to the uh Hurlingham club yes*

ICE_10 *oh right (0.950) so whe- how often do you play*

ICE_9 *I play uhm (0.220) once a week in the in the summer [...]*

ICE_10 *well mummy and I will have to come and watch you won't we*

ICE_9 *such fun*

ICE_10 *<laughing/> such fun (1.000) yes but what we h- what were we oh yes you saw Oliver Pemberton what did you do yesterday (EN-conv-02)*

In Example (1), the DM "but" is highlighting the contrast between a past and present situation: we see that the connection is very local and is further signaled linguistically by the repetition of the verb "to be" conjugated in different tenses; the two arguments in the scope of the DM are single adjacent utterances not separated by pauses. In Example (2), by contrast, several DMs (four "so" and one "but") are used by <ICE_10> to launch new higher-order discourse segments (often questions) which are themselves distributed across several turns. The "but" is particularly far-reaching since it closes the lengthy three-minute digression on Oliver Pemberton, his sister and bicycle polo and connects the final question of this extract ("what did you do yesterday") with the very first in the extract ("what did you do today"). The higher level of organization signaled by "but" is also reflected by the occurrence of word fragments and false-starts ("what we h- what were we"), which corroborates the link between major discourse boundaries and hesitations observed by the experimental studies mentioned above (e.g. Roberts and Kirsner 2000: 150).

This association is in fact telling of a hearer-oriented, strategic use of pauses and other performance phenomena which are not (only) the symptoms of trouble but can also perform signposting, forewarning functions (Clark and

Fox Tree 2002). The positive effects of both silent and filled pauses such as *uhm* have been the focus of many studies (e.g. Swerts 1998; Rendle Short 2004; Lundholm 2015) pointing in particular to their discourse-structuring function, similar to that of DMs. Therefore, the above-mentioned difference in processing load between microplanning and macroplanning might only concern the speaker's production efforts and not the interpretation effects for the hearer, where the coherence-building task might actually be reduced by explicit markers of discourse structure such as global-scope DMs and pauses.

The examples and references discussed in this section raise a number of hypotheses regarding potential cognitive correlates to the difference between local and global scope of DMs. Firstly, the functional similarity between DMs and pauses suggests that, when combined, these discourse-structuring signals might constitute reliable cues to a major boundary in the higher-order organization of talk. Concretely, the association between higher scope and co-occurrence of pauses will be tested on corpus data (see next section) to assess its reliability as an indirect cue to the variation in scope. Secondly, this first source of evidence will be refined by taking into account the position of the DM in relation to the turn (cf. the turn-initial uses of "so" in (2)) and to the dependency structure: this latter unit of reference allows to investigate the link between the scope of the DM and its degree of syntactic integration, mainly by comparing coordinating vs. subordinating conjunctions acting as DMs (e.g. *but* vs. *although*). Subordination is

hypothesized to correspond to DMs with a local scope, whereas global-scope DMs should be more attracted to "weak clause association" (Schourup 1999: 233), that is peripheral, syntactically non-integrated positions. Both syntax and co-occurrence with pauses are presently taken as indirect yet objective and operational cues to the variability of DM scope, assuming that they offer a more reliable methodological gateway to scope than the highly interpretative and potentially circular annotation of DMs arguments which might prove particularly challenging in spoken corpora.

3. *DisFrEn*: corpus and annotation

The role of syntax and pauses in DM scope will be tested on *DisFrEn*, a comparable English-French dataset where an inclusive, bottom-up selection of DMs has been annotated for positional and functional variables as well as co-occurrence with pauses and other hesitation phenomena. Space forbids to provide the full description of corpus design and annotation schemes (see [Author], 2017b) yet the major principles and criteria relevant to the present study will be laid out in this section. *DisFrEn* comprises around 15 hours of recordings and 161,700 words balanced across eight registers of English and French, including casual conversations, classroom lessons and political speeches. The dataset was compiled from several source corpora: most

English transcripts come from the British component of the International Corpus of English (ICE-GB, Nelson et al., 2002) while the French subcorpus is largely sampled from the VALIBEL dataset (Dister et al., 2009). Transcripts are audio-aligned and annotated under the EXMARaLDA software (Schmidt and Wörner, 2012)

In *DisFrEn*, discourse markers were identified onomasiologically (i.e. without a closed list), following a broad formal-functional definition (cf. the beginning of Section 2) operationalized after several phases of testing and identification experiments ([Author], 2015). In addition to the criteria of procedurality, syntactic optionality and high degree of grammaticalization (or fixation), a number of related devices were excluded from the DM category, such as filled pauses (*uhm*), tag questions (*isn't it*) or epistemic parentheticals (*I think*). The full list of annotated DMs amounts to more than 200 types and 8,743 tokens.

All identified DMs were annotated for several variables, of which four are of particular relevance to the present study. Each DM (including multi-word expressions such as *on the one hand*) was assigned a part-of-speech tag (henceforth POS) or "self-category", that is "the highest node in the tree which dominates the words in the connective but nothing else" (Pitler and Nenkova 2009: 14). Three types of position were then separately identified, taking as the reference unit either the turn (turn-initial, turn-medial, turn-final

or whole turn), the dependency structure (integrated vs. peripheral, left vs. right of the governing verb) or the clause (initial, medial, final).

Each DM was then functionally disambiguated according to a taxonomy of 30 senses (Table 1) grouped in four macro-functions or domains: this list is partly inspired by the PDTB 2.0 for discourse relations (e.g. cause) and González (2005) for speech-specific functions (e.g. monitoring).

Ideational	**Rhetorical**	**Sequential**	**Interpersonal**
cause	motivation	punctuation	monitoring
consequence	conclusion	opening boundary	face-saving
concession	opposition	closing boundary	disagreeing
contrast	specification	topic-resuming	agreeing
alternative	reformulation	topic-shifting	elliptical
condition	relevance	quoting	
temporal	emphasis	addition	
exception	comment	enumeration	
	approximation		

Table 1. List of functions grouped by domains

A random sample of 15% of the whole corpus was coded twice in order to assess intra-rater reliability: the agreement is substantial both for domains (Cohen's $\kappa = 0.779$) and functions ($\kappa = 0.74$). [Author] (in press) report lower scores for inter-rater reliability using this taxonomy: Fleiss' $\kappa = 0.563$ for domains and $\kappa = 0.406$ for functions, which can be explained by the very large number of values and the settings of the annotation experiment. To date, this taxonomy has been applied to speech and writing ([Author], 2015),

spoken Slovene (Dobrovoljc 2016), spoken Kinshasa Lingalá (Nzoimbengene 2016), gestures (Bolly 2015) and Belgian French Sign Language (Gabarro-López, forthc.).

In a last step of the analysis, all "disfluencies" (e.g. pauses, word fragments, repetitions) co-occurring with DMs were annotated, following [Author]'s (2016) multilingual typology. The present study will mainly focus on pauses (200ms or longer), either silent or filled (*uh*, *uhm*, French *euh*). Pause duration is not included in this analysis given that any threshold (for instance between "short" or "long" pauses) would require to take into account each speaker's average speaking rate, following Little et al. (2013). Configurations of DMs and pauses are detailed according to the position of each element within the cluster.

In sum, *DisFrEn* offers a relatively large, richly annotated dataset covering syntactic, functional and syntagmatic variables. Despite some methodological limitations (moderate replicability of the sense disambiguation task and absence of prosodic information), the annotated variables under scrutiny in this paper, viz. syntax and co-occurring pauses, are objective and reliable enough to ensure robust analyses of DM scope.

4. Syntax and pauses as indirect measures of DM scope

The following analyses test the extent to which position, degree of syntactic integration and co-occurrence with pauses can be used as reliable indirect cues to the divide between local and global scope of DMs. They target three pairs of DMs, each representing a different level of granularity: comparing two functions (Section 4.1), two syntactic classes (Section 4.2) and two uses of the same DM (Section 4.3). These pairs were selected for their intrinsic connection to varying degrees of scope: specific hypotheses will be laid out at the beginning of each subsection.

4.1 Function-specific: topic-shift vs. topic-resume

The first pair of DMs potentially associated with different degrees of scope concerns the *topic-shift* and *topic-resume* functions, respectively defined as i) a change of topic within or between turns carrying no or little connection with the previous context (including new subtopics) and ii) a return to a previous topic after a digression or a non-relevant segment. In terms of scope, topic-shift and topic-resume can be distinguished by the type of discourse unit that they introduce (new topic segment vs. regular utterance subordinated to an existing topic segment) and the typical distance between the related units (adjacent topics vs. utterances separated by a digression of varying length). The expectations are therefore not straight-forward: hierarchically, topic-shifts target higher-level discourse structure (global scope) yet the topic

segments themselves are adjacent (local scope), while topic-resuming DMs do not signal a major discourse boundary (local scope) but connect more or less distant units within a topic segment (global scope).

In *DisFrEn*, 234 occurrences of topic-resuming DMs and 291 topic-shifting DMs were annotatedited Looking at their syntactic position is not particularly interesting since both functions overwhelmingly favor the peripheral (i.e. not integrated) initial position in 88% and 92% of their occurrences, respectively. Position in the turn, however, reveals a strong, statistically significant difference between topic-shift and topic-resume as to the proportion of turn-initial uses: 32.3% of topic-shifting DMs are turn-initial against only 7.69% for topic-resuming ($z = -6.842$, $p < 0.001$)[4]. This result points to the specialization of topic-shifting DMs at a higher level of discourse organization, managing hierarchically larger units (i.e. whole turns).

This first positional cue to a more global scope of the topic-shift function is, however, not confirmed by co-occurring pauses, where we can observe a similar preference for the [pause+DM] pattern in 44% and 52% of turn-medial topic-resuming and topic-shifting DMs, respectively (turn-initial and turn-final DMs were excluded from this analysis since they are, by definition, less prone to co-occurring with pauses). This frequent co-occurrence with

[4] The z-ratio is used to test the significance of the difference between two independent proportions.

pauses in around half of all occurrences points to the discourse-structuring role of these DMs, compatible with the expectation of their global scope. Nevertheless, in both functions, a substantial proportion (around one quarter) of the turn-medial DMs occur in isolation. The two most frequent patterns (pause+DM and DM alone) are exemplified below.

(3) *I think she actually likes it but (0.727) she has a sense of proportion hold on here's a napkin oops (0.280)* by the way *did I mention my dustbin's been blown over in my back garden again (EN-conv-04)*

(4) [current lecturer of acoustics talking about how the acoustics class used to be done and his former classmate Jane]
she was actually *taking it for credit and it was a whole unit (0.420) so poor old little Janey (0.227) we were having a discussion with Bob actually about the uh the organization of the course […] Dick's written on […] what do the students think of the course (EN-conv-06)*

Moreover, the data shows that only a few of all tokens (64/407) co-occur with a filled pause (e.g. *uhm*), against our expectation of the link between global scope and the discourse-structuring function of filled pauses. The relatively

high proportion of isolated uses (as in Example (4)) and the low co-occurrence with filled pauses both tend to qualify the assumption that topic-shift and topic-resume systematically function globally and suggest that they might also be used locally. Another interpretation of this result suggests that the absence of (filled) pauses is not a systematic sign of local scope but might rather indicate a high degree of planning (planned speech) or a high level of interactional pressure on speakers not to lose the floor (interactive speech).

Still, the majority of cases expresses a rather far-reaching scope, as shown by the very low frequency of syntactically integrated DMs: 17/234 topic-resuming and 10/291 topic-shifting DMs occur within governed elements, as in Example (5) where the topic-shifting "then" is inserted before a complement ("in the name").

(5) [talking about the name of a company called "Ducks and Drake"]

BB_3 *Sir Francis Drake was based here […] and led his ships out to fight them*

BB_1 ok (0.560) and (0.220) what's the importance of the 'ducks' then *in the name*

BB_3 the 'ducks' are the specialist vehicles we use (EN-intf-02)

In sum, the two functions appear to act globally in their own distinct way (hierarchical structure vs. distance), which shows that a single measure of DM

scope might not be enough: the combination of syntax and pauses offers a more fine-grained picture yet does not suffice to oppose the degrees of scope between topic-shift and topic-resume. This first pair was therefore inconclusive and suggests to take a different, more formal approach to DM scope, namely to start from forms instead of functions, as we will now come to see.

4.2 POS-specific: subordination vs. coordination

The syntactic mechanisms of coordination (or parataxis) vs. subordination (or hypotaxis) have been widely studied, including in relation to DMs (cf. Pawley and Syder, 2000 on "clause-chaining" vs. "clause-integrating"; Castellà 2004; Blühdorn 2008). Coordinating conjunctions (henceforth CC) are very often used as DMs and constitute the most frequent members of the category (especially *and*, *but* and *so* in *DisFrEn*), while subordinating conjunctions (henceforth SC) such as *because*, *if* or *although* are also quite frequent, especially in formal monologues. Given that SC are syntactically governed and depend on a main verb, they are presently expected to function locally (i.e. take scope over single and adjacent utterances), in comparison with CC whose syntactic independence should be reflected by an attraction to peripheral positions and to pauses.

The data strongly confirms these expectations: 60% of all SC occur in integrated positions to the right of the governing verb (typically *although*) while the other 40% occur to its left (typically *if*); CC, on the other hand, largely prefer the initial non-integrated slot in 93% of the cases, with a few anecdotal occurrences in final position (cf. Mulder and Thompson, 2006 on final *but*) and left- or right-integrated positions, as in Example (6).

(6) *you can break into the pears if you want to* or *have a piece of choccy you've had plenty of veggies (EN-conv-01)*

These strong syntactic associations are to be expected from the rather circular definition of SC as syntactically integrated, although the positional behavior of CC is not as restrictededited Co-occurrence with pauses offers a more independent and interesting cue to the variation in scope: CC (restricted to turn-medial DMs) show no preference between isolated (DM alone) and co-occurring (pause+DM) contexts (40% vs. 39%) while SC are strongly attracted to the isolated uses in two thirds of all occurrences, against only 23% of co-occurrence. These findings tend to confirm the hypothesis of the larger scope of CC compared to SC, which can be related to their difference in syntactic integration.

Such an approach to different grammatical classes acting as DMs still covers a lot of variation, and it might be the case that syntactic and

syntagmatic behaviors within one class differ depending on specific functions or even particular DM expressions, which motivates the more fine-grained level of analysis in the next section.

4.3 DM-specific: so *expressing consequence vs. conclusion*

The last pair under investigation consists of two uses of the same DM, namely *so* expressing a consequence or a conclusion: these two functions share a semantic core (Arg2 is the result of Arg1), although in the former the relation is semantic or "objective" while in the latter the relation is pragmatic or "subjective" (see Pander Maat and Sanders 2000, 2001 on these notions). The epistemic distance involved in subjective relations such as *conclusion* could be related to a more global scope, acting on the mental representation of discourse rather than the local chaining of facts (as in *consequence* relations), which is expected to be reflected in the co-occurrence with pauses (more frequent for conclusive than consecutive *so*).

Indeed, only 29% of the 168 conclusive *so* occur in isolation against 61% of the 122 uses as *consequence*, while the [pause+DM] pattern represents 43% and 27% of their respective occurrences. Conclusive *so* is also quite frequent with a pause to its right [DM+pause] and at both sides [pause+DM+pause]. The most frequent patterns for each function are illustrated below.

(7) *from there we make our way round the citadel […] from there we then go down back to the start point (1.050)* so *it's a an all-encompassing tour covering all (0.227) ages of h- history of Plymouth (EN-intf-02)*

(8) *if I go home to visit say you will (0.240) notice when I come back (0.380) that I'm speaking with a Liverpool accent because my family do […] and it's around me on the Wirral* so *I come back talking* a *little bit more like a Liverpudlian (EN-intf-03)*

These tendencies are also observed in the French data with *donc* and tend to confirm the higher scope of subjective functions of DMs. However, they do not systematically apply to all objective-subjective pairs of relations: for instance, *because* and *if* are always more isolated than co-occurring with pauses regardless of their function, which might be explained by our previous finding on subordinating conjunctions and their preference for isolation.

5. Summary and discussion

This study revealed interesting patterns of position and co-occurrence with pauses which illustrate the potential of indirect yet operational cues to access

the multi-faceted notion of DM scope. In particular, high degree of syntactic integration and absence of co-occurring pauses was shown to be often associated with local scope, while DMs expressing a more global scope tend to occur outside the syntactic dependency structure, co-occur with pauses and introduce hierarchically larger and/or distant units.

The paper only provides a partial view of the phenomenon of local vs. global scope of DMs and even suggests that there might be more than one type of global scope (cf. Section 4.1). The notion requires more research from various frameworks: for instance, a constructionist approach to DMs (Fried and Östman 2005; Fischer 2010; [Author] 2017c) could further our understanding of the variation in scope by uncovering regular patterns of forms (syntactic class and position) and meanings (specific functions in context). Experimental paradigms should then confirm whether these discursive constructions are used and perceived by conversation participants as relevant units of cognitive processing (e.g. [pause+*so*] triggers the expectation of a global-scope relation).

Another promising research avenue is to dig further into the mapping between discourse segmentation, functional analysis and co-occurrence with pauses in order to converge multiple types of evidence for semantic-pragmatic phenomena. However, it is important that all levels of analysis remain independent from each other in order to avoid circularity, as opposed to existing models of spoken discourse segmentation (cf. RST, Val.Es.Co)

where either the relation and its arguments or the type of unit and its function are strongly inter-dependent. An indirect approach to scope, as presently illustrated, might be more methodologically robust and uncover constructions which are not only descriptively adequate but also "psychologically plausible", as advocated by the programme of cognitive pragmatics (Schmid 2012: 4-5).

Analyzing DM scope, whether directly through full-text segmentation or indirectly through converging formal and functional evidence, always involves some subjectivity on the linguist's part and raises the issue of how far off-line annotations can go without putting words in the speaker's mouth: if functional annotation of DMs is a complex undertaking (e.g. Spooren and Degand, 2010), should we strive to add systematic argument identification on top of it? Is sense disambiguation already too subjective and interpretative to be reliable? According to Glynn (2010), there are ways to operationalize the analysis (e.g. documenting guidelines, inter-rater agreement) and converging evidence through statistical modelling of independent variables is strongly encouraged as a growing method for corpus-driven cognitive semantics: "confirmatory techniques, based entirely on highly subjective annotation, not only produce coherent results but results that can accurately predict the data" (Glynn 2010: 260). The exact balance between objectivity and subjectivity, quantitative and qualitative, top-down (direct) and bottom-up (indirect) is yet to be found and the present paper only paves the way for a critical

reconsideration of existing approaches to DM scope and, more generally, of the inter-dependence between annotation variables, focusing in particular on the interface between syntax and discourse.

References

[Author] 2015. "Using a unified taxonomy to annotate discourse markers in speech and writing." In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11), April 14th, London, UK,* ed. by Harry Bunt: 14–22.

[Author] 2016. "Annotation manual of fluency and disfluency markers in multilingual, multimodal, native and learner corpora. Version 2.0." Technical report, Université catholique de Louvain and Université de Namur.

[Author] 2017a. "Towards an operational category of discourse markers: A definition and its model." In *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*, ed. by Chiary Fedriani, and Andrea Sansó, 99-124. Amsterdam: John Benjamins.

[Author] 2017b. "Discourse markers and (dis)fluency in *DisFrEn*: Variation and combination in English and French." *International Journal of Corpus Linguistics* 22 (2): 242-269.

[Author] 2017c. “From co-occurrence to constructions: patterns of discourse markers and disfluencies across registers in English and French.” Paper given at the 14th International Cognitive Linguistics Conference (ICLC-14), July 10–14, Tartu, Estonia.

[Author] in press. “Discourse markers in speech: Characteristics and challenges for corpus annotation.” *Dialogue and Discourse*.

[Author] in press. “Reliability vs. granularity in discourse annotation: What is the trade-off?” *Corpus Linguistics and Linguistic Theory*.

Baldridge, Jason, and Alex Lascarides. 2005. “Annotating Discourse Structure for Robust Semantic Interpretation.” In *Proceedings of the 6th International Workshop on Computational Semantics*.

Beattie, Geoff. 1980. “Encoding Units in Spontaneous Speech: Some Implications for the Dynamics of Conversation.” In *Temporal Variables in Speech*, edited by Hans-Wilhelm Dechert, and Manfred Raupach, 131–143. Den Hague: Mouton.

Blühdorn, Hardarik. 2008. “Subordination and Coordination in Syntax, Semantics and Discourse: Evidence from the Study of Connectives.” In *“Subordination” versus “Coordination” in Sentence and Text*, edited by Catherine Fabricius-Hansen, and Wiebke Ramm, 59–85. Amsterdam: John Benjamins.

Bolly, Catherine. 2015. “Towards Pragmatic Gestures: From Repetition to Construction in Multimodal Pragmatics.” Paper Given at the 13th

International Cognitive Linguistics Conference (ICLC-13), July 20–25, Newcastle, UK.

Bunt, Harry. 2012. "Multifunctionality in Dialogue." *Computer Speech and Language* 25: 222–245.

Cabedo, Adrián, and Salvador Pons (eds). 2013. *Corpus Val.Es.Co*. [online: http://www.valesco.es].

Castellà, Josep. 2004. *Oralitat i Escriptura. Dues Cares de la Complexitat del Llenguatge*. Barcelona: Publicacions de l'Abadia de Montserrat.

Clark, Herbert H., and Jean E. Fox Tree. 2002. "Using *uh* and *um* in Spontaneous Speaking." *Cognition* 84: 73–111.

Cuenca, Maria Josep. 2013. "The Fuzzy Boundaries between Discourse Marking and Modal Marking." In *Discourse Markers and Modal Particles. Categorization and description* [Pragmatics and Beyond New Series 234], edited by Liesbeth Degand, Bert Cornillie, and Paola Pietrandrea: 191–216. Amsterdam: John Benjamins.

Das, Debopam, Maite Taboada, and Paul McFetridge. 2015. *RST Signalling Corpus LDC2015T10*. Philadelphia: Linguistic Data Consortium.

Deppermann, Arnuld, and Susanne Günthner. 2015. *Temporality in Interaction*. Amsterdam: John Benjamins.

Dister, Anne, Michel Francard, Philippe Hambye, and Anne-Catherine Simon. 2009. "Du corpus à la banque de données. Du son, des textes et

des métadonnées. L'évolution de la banque de données textuelles orales VALIBEL (1989-2009)." *Cahiers de Linguistique* 33 (2): 113–129.

Dobrovoljc, Kaja. 2016. "Annotation of Multi-word Discourse Markers in Spoken Slovene." Poster given at *Discourse Relational Devices (LPTS 2016)*, January 24–26, Valencia, Spain.

Estellés, Maria, and Salvador Pons. 2014. "Absolute Initial Position." In *Discourse Segmentation in Romance Languages*, edited by Salvador Pons: 121–155. Amsterdam: John Benjamins.

Fischer, Kerstin. 2000. *From Cognitive Semantics to Lexical Pragmatics*. Berlin: Mouton de Gruyter.

Fischer, Kerstin. 2010. "Beyond the Sentence: Constructions, Frames and Spoken Interaction." *Constructions and Frames* 2 (2): 185–207.

Fried, Mirjam, and Jan-Ola Östman. 2005. "Construction Grammar and Spoken Language: The Case of Pragmatic Particles." *Journal of Pragmatics* 37: 1752–1778.

Gabarró-López, Sílvia. Forthcoming. "Marqueurs du discours en langue des signes de Belgique francophone (LSFB) et langue des signes catalane (LSC): les 'balise-listes' et les 'palm-ups'." In *Marcadores del Discurso y Lingüística Contrastiva en las Lenguas Románicas*, edited by Óscar Loureda, Giovanni Parodi, Martha Rudka, and Shima Salameh. Madrid: Iberoamericana Vervuert.

Glynn, Dylan. 2010. “Testing the Hypothesis. Objectivity and Verification in Usage-based Cognitive Semantics.” In *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches* [Cognitive Linguistic Research 46], edited by Dylan Glynn, and Kerstin Fischer: 239–269. Berlin: De Gruyter Mouton.

González, Montserrat. 2005. “Pragmatic Markers and Discourse Coherence Relations in English and Catalan Oral Narrative.” *Discourse Studies* 77 (1): 53–86.

Greene, John, and Joseph N. Cappella. 1986. “Cognition and Talk: The Relationship of Semantic Units to Temporal Patterns of Fluency in Spontaneous Speech.” *Language and Speech* 29 (2): 141–157.

Hansen, Maj-Britt Mosegaard. 2006. “A Dynamic Polysemy Approach to the Lexical Semantics of Discourse Markers (with an exemplary analysis of French *toujours*).” In *Approaches to Discourse Particles*, edited by Kerstin Fischer: 21–41. Amsterdam: Elsevier.

Lenk, Uta. 1998. “Discourse Markers and Global Coherence in Conversation.” *Journal of Pragmatics* 30: 245–257.

Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Little, Daniel R., Raoul Oehmen, John Dunn, Kathryn Hird, and Kim Kirsner. 2013. “Fluency Profiling System: An Automated System for Analyzing

the Temporal Properties of Speech." *Behavioral Research Methods* 45 (1): 191–202.

Lundholm, Kristina. 2015. *Production and Perception of Pauses in Speech*. Doctoral dissertation, University of Gothenburg.

Mann, William, and Sandra Thompson. 1988. "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization." *Text* 8 (3): 243–281.

Mulder, Jean, and Sandra Thompson. 2006. "The Grammaticalization of *but* as a Final Particle in English Conversation." In *Selected Papers from the 2005 Conference of the Australia Linguistic Society*, Edited by Keith Allan.

Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Nzoimbengene, Philippe. 2016. *Les 'Discourse Markers' en Lingála. Étude Sémantique et Pragmatique sur Base d'un Corpus de Lingála de Kinshasa*- Doctoral dissertation, Université catholique de Louvain.

Pander Maat, Henk, and Ted J. M. Sanders. 2000. "Domains of Use and Subjectivity. On the Distribution of three Dutch Causal Connectives." In *Cause, Condition, Concession and Contrast: Cognitive and Discourse Perspectives*, edited by Bernd Kortmann, and Elizabeth Couper-Kuhlen: 57–82. Berlin: Mouton de Gruyter.

Pander Maat, Henk, and Ted J. M. Sanders. 2001. “Subjectivity in Causal Connectives: An Empirical Study of Language in Use.” *Cognitive Linguistics* 12 (3): 247–273.

Pawley, Andrew, and Frances H. Syder. 2000. “The One-clause-at-a-time Hypothesis.” In *Perspectives on Fluency*, edited by Heidi Riggenbach: 163–199. Ann Arbor: The University of Michigan Press.

Pitler, Emily, and Ani Nenkova. 2009. “Using syntax to disambiguate explicit discourse connectives in text.” In *Proceedings of the ACL-IJCNLP Conference Short Papers*: 13–16.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. “The Penn Discourse Treebank 2.0.” In *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), May, Marrakech, Morocco*.

Rendle-Short, Johanna. 2004. “Showing structure: using um in the academic seminar.” *Pragmatics* 14 (4): 479–498.

Roberts, Benjamin, and Kim Kirsner. 2000. “Temporal cycles in speech production.” *Language and Cognitive Processes* 15 (2): 129–157.

Sanders, Ted J. M., Wilbert Spooren, and Leo Noordman. 1992. “Toward a taxonomy of coherence relations.” *Discourse Processes* 15: 1–35.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schmid, Hans-Jörg. 2012. “Generalizing the apparently ungeneralizable. Basic ingredients of a cognitive-pragmatic approach to the construal of meaning-in-context.” In *Cognitive Pragmatics*, edited by Hans-Jörg Schmid: 3–22. Berlin: De Gruyter.

Schmidt, Thomas, and Kai Wörner. 2012. „EXMARaLDA“. In *Handbook on Corpus Phonology*, edited by Jacques Durand, Gut Ulrike, and Gjert Kristoffersen: 402–419. Oxford: Oxford University Press.

Schourup, Lawrence. 1999. “Discourse markers: A tutorial”. *Lingua* 107: 227–265.

Spooren, Wilbert, and Liesbeth Degand. 2010. “Coding coherence relations: reliability and validity.” *Corpus Linguistics and Linguistic Theory* 6 (2): 241–266.

Stent, Amanda. 2000. “Rhetorical structure in dialog.” In *Proceedings of the 2nd International Natural Language Generation Conference (INLG’2000)*.

Swerts, Marc. 1998. “Filled pauses as markers of discourse structure.” *Journal of Pragmatics* 30: 485–496.

Unger, Christoph. 1996. “The scope of discourse connectives: implications for discourse organization.” *Journal of Linguistics* 32: 403–438.

Webber, Bonnie, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. “Discourse annotation of conjoined VPs.” In *Conference Handbook of the 2nd Action Conference of TextLink, April, Budapest, Hungary*: 135–140.